Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 621.313
DOI: https://doi.org/10.24867/18BE07Josipovic

SKLADIŠTENJE PODATAKA U DNK MOLEKULIMA

DNA-BASED DATA STORAGE

Andrea Josipović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTREHNIKA I RAČUNARSTVO

Kratak sadržaj – *Potrebe za novim alternativama skladištenja sve veće količine podataka su evidentne. Čuvanje podataka u DNK molekulima se javlja kao jedna od opcija koja ima brojne prednosti u odnosu na postojeće mogućnosti. Kakav uticaj može imati na čitavu sferu skladištenja podataka, način realizacije i njene prednosti i mane detaljno su analizirane u ovom radu.*

Ključne reči: *DNK molekul, skladištenje podataka*

Abstract – *The necessity of new alternatives for the purposes of digital data storage is greatly increasing. Storing data in DNA molecules is one of the options which has many benefits in comparison to already existing storage mediums. How great of an impact can it have to the whole field of data storage, it's realizations and (dis)advantages are thoroughly analyzed in this paper.*

Keywords: *DNA molecule, data storage*

1. UVOD

Ljudi imaju prirodnu potrebu za novim saznanjima koja čuvaju i akumuliraju za neke buduće primene. Upravo to je karakteristika nasleđena intelektualnom evolucijom čoveka. Dezoksiribonukleinska kiselina (DNK) omogućava upravo pomenutu sposobnost čuvanja za potencijalnu buduću upotrebu, prvenstveno zbog osobina izdržljivosti i kompaktnosti koje je odlikuju, kao i zbog načina skladištenja sličnog dosadašnjim kompjuterskim sistemima. Oblast DNK skladištenja koja je u razvoju ima potencijala da transformiše naučnu fantastiku u realnost pomoću "uređaja" koji može da nam stane u dlan, a da pritom u sebi sadrži sve globalne podatke generisane na godišnjem nivou.

2. IZAZOVI TRADICIONALNIH MEDIJUMA

Globalno informaciono doba karakterisano je stvaranjem, kupovanjem, prodavanjem i nagomilavanjem podataka u toj meri da nadmašuje ljudske mogućnosti analiziranja, skladištenja i čuvanja. U tom smislu velika količina podataka predstavlja veliki problem.

Na osnovu istraživanja IDC Global DataSphere kompanije, za podatke koji su generisani na svetskom nivou (podrazumevajući nove i kopije starih) predviđa se da će rasti za 23% na godišnjem nivou u periodu od 2020. do 2025. godine.
Prema toj stopi, zaključno sa 2025. godinom, ukupna količina podataka dostići će 180 zetabajta. Udeo porasta ne čine samo nove informacije koje je potrebno sačuvati, već i replike postojećih. Glavni razlog velikog broja kopija jeste Cloud online mogućnost skladištenja, gde se podaci višestruko multipliciraju kako bi se smanjila verovatnoća oštećenja informacije u bilo kakvom smislu i obezbedilo brže dolaženje do podataka jer je time preskočen korak oporavka podatka do prvobitnog stanja. Masovno čuvanje sa sobom nosi i ogromna materijalna ulaganja, koja nisu dostupna širokim narodnim masama.

Današnji skladišni medijumi (magnetni, poluprovodnički i drugi) mogu sa odgovarajućim održavanjem da čuvaju podatke i do nekoliko decenija. Ipak, oni imaju ograničen životni vek i njihove mogućnosti vremenom degradiraju. Sve pomenuto za posledicu ostavlja neophodnost periodične provere i kontinuirani monitoring kako bi se očuvao integritet podataka koje čuvaju. Dodatna otežavajuća okolnost jeste promenljiv format sadržaja koji se menja iz generacije u generaciju jer se i procesi čitanja i upisivanja sa razvojem menjaju. Mana na koju ukazuju pojedine estimacije jeste činjenica da će skladišni centri koji su 2018. godine trošili 1% totalne energije na globalnom nivou, trošiti 3 ili 4 puta više u narednoj deceniji. Za zetabajtsku skalu koja je savremeni cilj skladištenja ovi trendovi predstavljaju ozbiljne prepreke kako u materijalnom, tako i u praktičnom smislu. Bolje rečeno, sve ukazuje na potrebu nekog novog pristupa u hijerarhiji skladištenja koga će karakterisati materijalna isplativost i kapaciteti koji će moći da proprate eksponencijalni rast podataka u 21. veku [1][2].

3. DNK MOLEKUL KAO SKLADIŠNI MEDIJUM

DNK je primarni genetički materijal. Osnovni je nosilac genetičke informacije (gena) kod živih bića, sa izuzetkom pojedinih virusa kod kojih tu ulogu ima ribonukleinska kiselina (RNK). Njegova struktura je predstavljena polinukleotidnim lancem koji se sastoji od niza nukleotida. Sami nukleotidi su složene strukture koju čine azotna baza, pentozni šećer i fosfatna grupa. Govoreći o strukturi nukleotida u DNK, azotna baza je diferencirajući faktor. Ona se može pojaviti u nekoj od četiri varijante, dok su preostala dva elementa – pentozni šećer dezoksiriboza i fosfatna grupa – uvek isti. Varijante azotnih baza su:

- Adenin – A
- Guanin – G
- Timin – T
- Citozin – C

Kostur polinukleotidnog lanca čine veze uzastopnih susednih dezoksiriboza i fosfatnih grupa. Unutrašnjost polinukleotidnog lanca konstruišu veze komplementarnih azotnih baza pomoću vodoničnih veza: A-T (pirini) i G-C (piramidini), rezultujući u četiri moguće varijacije. Dva

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio doc. dr Mladen Kovačević.

paralelna polinukleotidna lanca se međusobno uvijaju oko zamišljene zajedničke ose u dvojnu spiralu formirajući dvostruki heliks (dvostruka spirala) [2].

3.1 Prednosti primene DNK molekula

Priroda je u svoje svrhe razvila molekule koji imaju neverovatne kapacitete skladištenja. Jedan od tih molekula jeste i DNK, molekularni depozit biološke informacije. Gram DNK koji se sastoji od 10^{21} DNK baze može da čuva 10^8 terabajta binarnih podataka (nula i jedinica). Volumetrijski pristup fizičkog što se izvrši sinteza, DNK se trajno čuva u kapsuli i time štiti od spoljašnjih uticaja. Postoje mnogi načini na koji se ovo može uraditi, podrazumevajući da se prilikom njihovog postavljanja u kapsulu ubrizgava i inertni gas ili neki drugi hemijski materijal koji podržava njegovo očuvanje. Svaka opcija čuvanja DNK mora koristiti neki unapred izabrani materijal kapsule u kojem će se čuvati. Postoji više mogućnosti skladištenja sa svojim prednostima i manama koje utiču u odabiru metode, u zavisnosti od ograničenja i prednosti koje svaka od njih nosi.

1. **Izdvajanje DNK iz biblioteke –** Kada postoji potreba da se dođe do podatka koji je skladišten u DNK, to podrazumeva da se iz čitave biblioteke DNK koji čuvaju informacije izdvoji traženi i potom pripremi za sekvenciranje. Vrlo često ovaj korak podrazumeva i pravljenje kopija molekula za različite kasnije potrebe. Kako bi se izbeglo čitanje cele biblioteke da bi se došlo do željenog DNK molekula primenjuje se random pristup. On omogućava da se pronađe traženi podatak za čitanje bez potrebe da se prethodno prođe kroz celu biblioteku. Ova realizacija je mnogo jednostavnija u tradicionalnim skladišnim medijumima, dok je na molekularnom nivou izazovnije zbog nedostatka fizičke organizacije podataka u DNK. Jedan primer random pristupa jeste PCR (engl. polymerase chain reaction) metoda.

2. **Sekvenciranje (čitanje) –** DNK sekvenciranje je pojam koji opisuje veliki broj tehnika za detekciju redosleda baza u polinukleotidnim lancima. Ovaj korak je među prvima istraživan kada se skladištenje DNK pominje. Među prvim pokušajima DNK skladištenja se pominje metoda čitanja - Lančano završavanje (engl. chain termination sequencing). Njen tvorac Sanger njome je napravio veliku prekretnicu u ovoj oblasti i prvi put dokazao da je čitanje iz DNK izvodljivo. Od 1990-ih je aktuelna tehnologija sekvenciranja zvana Nova generacija sekvenciranja (engl. Next generation sequencing - NGS) koja je svojim karakteristikama značajno poboljšala ovaj korak. Koristi mogućnost paralelizacije kako bi povećala skalabilnost, brzinu i kvalitet prenosa podataka. NGS predstavlja čitavu skupinu različitih metoda sekvenciranja od kojih su danas dve u širokoj komercijalnoj upotrebi [5][7].

Tabela 1. *Tabelarno poređenje DNK kapaciteta sa dosadašnjim pristupima na osnovu nekoliko karakteristika [3].*

Device	Data retention	Storage density	Power usage (watts/gigabyte)	Access time
Hard disk	10 years	$\sim 10^{13}$	~0.04	7 ms
Flash memory	~10 years	$\sim 10^{16}$	~0.01–0.04	5 ns
DRAM	~64 ms or less	$\sim 10^{13}$	A few tenths of watt	60 ns
Cellular DNA	> 100 years	$\sim 10^{19}$	$< 10^{-10}$	Slower than conventional media

5. DNK „SKRIVENO“ UPISIVANJE

Tajno upisivanje se koristi kako bi se sprečio pristup informacijama nedozvoljenim licima i njihovo nedozvoljeno korišćenje. Za realizaciju ovih ciljeva primenjuju se rešenja iz kriptografije i steganografije – naučnih disciplina koje su već vekovima aktuelne i čija se primena pronalazi u različitim oblastima.

5.1. DNK kriptografija

Mogućnost molekula DNK da skladišti, procesira i prenosi informacije je inspirisala ideju DNK kriptozaštite. Zasniva se na korišćenju četiri azotne baze – A, G, T i C za sva računanja. Glavna prednost DNK računanja jeste paralelizacija DNK molekula. Matematički aspekt u kriptografiji je zamenjen za hemijski, što onemogućava razbijanje ovog vida kriptografije konvencionalnim metodama ili kvantnim računanjem. Skladišni kapaciteti su još jedan razlog zbog kojeg je fokus na DNK kriptografiji. Jedan gram DNK molekula se sastoji od 10^{21} DNK baza koje mogu sadržati do 10^8 terabajta.

Gehlani i saradnici su prvi ispitivali mogućnosti kreiranja kriptosistema koristeći DNK molekule 1999. godine. Definisali su OTP šifrovanje (engl. One Time Pad) pomoću dve tehnike: DNK supstitucije i korišćenje XOR logičke operacije na nivou bita. Budući da OTP metoda šifrovanja podrazumeva i ključ kojim se definiše pravilo šifrovanja i koje jedino može dešifrovati istu tu poruku, u slučaju sa DNK dolazi se do ideje da lanac DNK ima ulogu ključa. Bitno je pomenuti da ne postoji jedan ključ kojim se šifruje svaki deo poruke, već postoji čitava biblioteka ključeva kojima je svako od šifrovanja definisano.

Kada se radi o metodi supstitucije koraci su sledeći:

1) Za polazni korak uzima se binarna poruka koja je predstavljena polinukleotidnim lancem DNK dužine n i koja je podeljena na manje delove fiksne dužine.

2) Na jedinstven i slučajan način se vrši mapiranje nad svakim manjim delom definisane dužine binarne poruke u šifrovanu.

3) Svaka ponavljajuća celina je na slučajan način postavljena na poziciju u nizu, potom izolovana i na kraju klonirana kako bi se formirala biblioteka ključeva.

4) Dolazi do hibridizacije, odnosno produžavanja ključa pomoću enzima polimeraze uz praćenje koraka protokola DNK replikacije [1][4].

5.2. DNK steganografija

DNK steganografija omogućava prenos poverljivih informacija u živim organizmima sve dok su PCR prajmeri i tajni ključ poznati svakom primaocu. Ponovo, Gehlani i saradnici su prvi predložili 1999. godine metod steganografije koji uključuje i DNK. Koraci su sledeći:

1) Polazni sadržaj predstavljen je preko polazne sekvence DNK lanaca.
2) Oni se potom taguju pomoću tajnog ključa koji je takođe u formi DNK.
3) Polazne sekvence se mešaju sa nasumično formiranim sekvencama DNK koje su označeni kao ometači.
4) Ako je tajni ključ poznat primaocu, tada polazne sekvence DNK lanaca mogu da se izdvoje iz pomešanih lanaca prateći korake protokola pročišćavanja po srodnosti. Jednostruka sekvenca koja se koristi je komplementarna sekvenca i predstavlja tajni ključ [6].

6. GREŠKE PRILIKOM ŠIFROVANJA I DEŠIFROVANJA PODATAKA

Sinteza i sekvenciranje su skloni greškama. Nekoliko istraživanja do sada je pokazalo da postoji prosečna stopa greške od 1%, takvih da taj procenat pozicija u DNK neće nakon upisivanja i kasnijeg čitanja sadržati istu informaciju u njoj. Za sada je sigurno da će procesi koji se primenjuju u ova dva koraka u manjoj ili većoj meri uticati na stopu greške i da PCR metoda i korak skladištenja mogu dovesti i do brisanja podataka. Ovakve performanse nije poželjno ponuditi korisnicima. Stoga, bitno je da se naknadno implementira korak ispravljanja grešaka. U ove svrhe, postoji cela oblast u okviru kompjuterskih nauka koja se zove teorija informacija ili teorija kodovanja koja se fokusira na obezbeđivanje nepromenjenog podatka pri preuzimanju iz memorije, uprkos svim smetnjama koje se javljaju. Za razliku od klasičnih skladišnih opcija, DNK može manifestovati i brisanje ili dodavanje baza i tako otežati pravu traženu informaciju pri čitanju [8][9].

6.1. Osnove ispravljanja grešaka

Svode se na dodavanje redundantnih informacija koje povećavaju verovatnoću da se originalna informacija na kraju preuzme u celosti i nepromenjena, čak i u prisustvu grešaka ili izbrisanih podataka. Što je redundatna stopa veća, to je veća tolerancija na greške (ili gubitke). Kada se govori o reduntantnosti, postoje dva osnovna tipa:

1. **Fizička redundantnost** – odnosi se na veliki broj kopija sekvenci DNK. DNK sinteza proizvodi veliki broj fizičkih kopija iste DNK sekvence.

2. **Logička redundantnost** – odnosi se na integrisanje dodatnih informacija kada se koduju biti u DNK sekvence.

Iako fizička redundantnost utiče na bolju toleranciju i eliminiše uticaj manje količine grešaka iz sinteze i sekvenciranja, to ipak nije dovoljno za garanciju skladištenja bez greške sa velikom pouzdanošću. Prve metode za redukciju greške koje su u primeni danas su se prvi put pojavile 1940. godine. Svima im je zajedničko što dodaju redundansu originalnom podatku pre nego što je sačuvaju ili transmituju kroz kanal. Prijemnici dodatu redundansu koriste da provere da li je primljeni sadržaj konzistentan i ako nije da se rekonstruiše izvorna informacija. Sama količina rendundanse koja se može dodati zavisi od šuma kom podaci mogu biti izloženi, metode koja se koristi i verovatnoće uspešnog dekodovanja koja želi da se postigne [8].

7. SADAŠNJOST I BUDUĆNOST

Sa sadašnje tačke gledišta, verovatno je da će vreme koje je potrebno da se izvrši korak čitanja i dalje biti visoko (nekoliko minuta do sati) u nekoj skorijoj budućnosti. Ipak, dokle god je količina podataka koja može istovremeno da se upisuje ili kasnije čita visoka (zahvaljujući paralelizaciji), in vitro DNK skladištenje može da bude pomoćno ili čak glavno sredstvo u komercijalnim svrhama. Iako predstoje ozbiljne prepreke koje treba da se premoste u materijalnom i drugim aspektima pomenutim u ovom odeljku, treba pomenuti da za potrebe DNK skladištenja tačnost može da se žrtvuje za brzinu izvršavanja, fizička redundansa se može značajno smanjiti. To sve omogućava dalji razvoj i poboljšanje performansi kako sinteze tako i sekvenciranja. Očekuje se da će se time troškovi smanjiti jer će biti prilagođeni većim podlogama sinteze i većim bečevima DNK. U korist tome ide i manji broj neophodnih kopija DNK sekvenci koji je potreban da se obezbedi pouzdanost čitavog sistema. Iako postoje dokazi koji ukazuju na mogućnost čitanja sadržaja iz nekoliko godina starih DNK molekula, ipak postoji mogućnost degradacije brže ili sporije u zavisnosti od uslova u kojima se nalazi. Visoke temperature, vlažnost i izloženost ultravioletnom svetlu su samo neki od njih. Danas i dalje najveći deo sinteze i sekvenciranja izvode ljudi u laboratorijskim uslovima. Biblioteke DNK treba da omoguće automatizaciju putanja do željene kapsule i skalabilnost bez značajnog ugrožavanja gustine.

DNK ima potencijala za višestruku primenu u različitim oblastima, kako danas tako i u budućnosti. Trenutno se radi na razvijanju kompjutera koji se zasnivaju na kvantnoj teoriji, teoriji koja se zasniva na modernoj fizici i objašnjava prirodu, ponašanje materije i energije na atomskom i podatomskom nivou. Kompanije kao što su Microsoft, IBM i Google ulažu velike količine novca upravo u takva istraživanja. Kada se govori o kvantnoj mehanici, ona ne dozvoljava da se podaci skladište direktno u neki medijum, nego moraju da se konvertuju u kvantne bite pre samog čina skladištenja. Zato, kada kvantni kompjuteri u budućnosti postanu komercijalno pristupačni, istraživači će jednostavno iskoristiti njihove funkcije i integrisati sa DNK. Pored kvantnih mašina ispituju se načini na koje se mogu konstruisati kognitivne mašine koje se zasnivaju na konvergenciji biološke i fizičke inteligencije, kao i biološki inspirisani roboti koji koriste DNK u svom funkcionisanju.

Za korak šifrovanja originalnog sadržaja može se iskoristiti DNK. Jednostruki i dvostruki lanci DNK se konkretno mogu koristiti za šifrovanje ne samo podataka, čak i čitavog softvera. Teži se ka tome da će sistem moći da semantički izanalizira neki šablon, na primer sliku ili zvuk, te da iz tog šablona može da izdvoji sve glavne karakteristike ili informacije iz njih [4][10].

8. ZAKLJUČAK

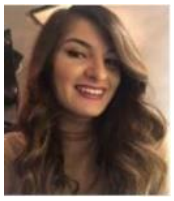
Mogućnost skladištenja podataka u DNK dokazana je kao izvodiva 80-ih godina prošlog veka i od tada su istraživanja na ovu temu stalno aktuelna, a napredak posledično uočljiv. Dalje u budućnosti, postizaće još bolje rezultate sa razvojem oblasti kompjuterskih nauka. To, između ostalog, podrazumeva uređaje za čitanje DNK koji su znatno brži od današnjih, random pristup koji će se usavršiti automatizacijom uz pomoć softvera koji su specijalno kreirani za njih.

Kako se tehnologije budu razvijale, tako će DNK skladištenje postati neizostavni deo šireg ekosistema novih kompjuterskih tehnologija koje se zasnivaju na sintezi sintetičke biologije i poluprovodničke industrije.

9. LITERATURA

[1] V. Demidov, "*DNA Beyond Genes*", Springer 2020.
[2] https://www.cs.utexas.edu/~bornholt/dnastorage-asplos16/
[3] P. Darshan, A. M. Kutubuddin, J. B. Mirza, S. Alaka., B. Deeptirekha, D. Manaswini, "*DNA as a digital information storage device: hope or hype*?", Springer 2018.
[4] Članovi DNA Data Storage alijanse, "*An introduction to DNA data storage*", DNA data storage alliance, 2021.
[5] "*The future of DNA data storage*", Potomac Institute for Policy Studies, 2018.
[6] J. E. Lauzan, J. Hall, M. Smith, "*Could DNA be the next big thing in data storage?*", Atos Scientific Community, 2015
[7] "*DNA-based digital storage*", Twist Biosence, 2017.
[8] P.Y. De Silva and G. U. Ganegoda, "*New Trends of Digital Data Storage in DNA*", BioMed research International, 2016.
[9] B. Carmean, L. Ceze, G. Seelig, K. Stewart, K. Strauss, M. Willsey, "*DNA Data Storage and Hybrid Molecular–Electronic Computing*", IEEE, 2018.
[10] M. Mondal, K. S. Ray, "*Review on DNA Cryptography*", International Journal of Information & Computation Technology, 2019.
[11] S. Namasudra, G. C. Deka, "*Advances of DNA computing in cryptography*", Taylor & Francis Group, 2019.

Kratka biografija:



Andrea Josipović rođena je u Novom Sadu 1997. godine. Osnovne studije na Fakultetu Tehničkih nauka završila je 2020. godine. Master rad na istom fakultetu iz oblasti Elektrotehnike i računarstva – Obrada signala odbranila je 2021. godine.

kontakt: andrea.josipovich@gmail.com